ChinaXiv合作期刊 图书情报工作
第66卷第7期 2022年4月

# 故事线构建及可视化、情感化、场景化应用探索*
## ——以《张学良口述历史》为例

王阮 邓君
吉林大学商学与管理学院 长春 130012

摘 要：[目的/意义]以故事线梳理史实脉络，不仅对于描绘与把握历史发展方向具有一定的理论和现实指导意义，同时也为人文领域的知识发现提供新的技术实现方式和创新性研究视角。[方法/过程]提供一种基于文本数据的故事线构建及可视化、情感化、场景化的研究范式，通过爬梳《张学良口述历史》文本作为数据源进行实证探索。采用jieba分词对《张学良口述历史》初始数据源进行数据清洗获取实验数据源，应用LDA主题模型获取主题分布并进行t-SNE数据降维呈现主题模块，借助SnowNLP情感词典挖掘情感特征词，梳理张学良情感演化阶段，进行故事线构建。[结果/结论]通过构建张学良故事线，实现人物、地点、事件、时间、情感等多维要素的动态互促。

关键词：故事线 可视化 情感化 场景化 张学良
分类号：G254
DOI：10.13266/j.issn.0252-3116.2022.07.002



## 1 引言

伴随数字人文迅猛发展，文本处理和图像分析能力革新，数据类型和规模呈指数增长趋势。在海量数据面前，用户不再囿于数量堆栈，也不再满足于文本内容、文本关系挖掘，而是致力于提高文本数据处理的效率和深度[1]，探寻多元复杂的数据规律及数据间潜藏的语义关系，快速发现核心内容和潜在知识关联。

目前，有关历史文化的研究多趋向数字化及数据构建[2]，鲜从故事线视角探略。作为一种新兴叙事手段，故事线是为了提升数据的可理解性、可记忆性及可体验性，将“数据”还原或关联至特定情景[3]，基于基础时间线隐喻的视觉表达，以事件作为记忆的存储单元[4]，能直观、交互式展现数据信息、深度解读数据，本质是以“故事叙述”方式呈现“从数据中发现的洞察”，可以还原情景、移植情景和虚构情景[5]。

“数智”时代，人文学者面临着使用传统手段感知数据的技术困境和进行数据交互认知的人文迷思，亟需寻求兼具技术与人文框架的信息交流与传播工具[6]。而故事线凸显了人文关怀与数字技术的交互渗透[7]，并且与数字人文多元化、融合化、可拓展性的特征深度契合，对于人文学者躲避海量低价值密度信息淹没，摆脱数据感知与交互困境具有重要的工具价值，成为数字人文方法论甚至技术体系的重要环节[6]。因此，本文从数字人文视角出发，择取《张学良口述历史》文本为数据源，并以此为基点构建张学良故事线，不拘泥于人—事—地—时等基础关联信息的呈现，还注入了情感演化时期，使故事线表现形式更加饱满、丰富，不仅突破了传统知识组织的二维平面空间，实现了多模态（可视化、情感化与场景化）知识组织聚合、知识关联拓展与多维知识耦合，而且对于描绘、把握与分析历史人物发展走向，更好地理解、梳理与挖掘要素具有重要启示意义。同时，也为人文领域学者研究提供了一种基于文本数据的故事线构建及可视化、情感化、场景化的创新型研究范式。

## 2 文献回顾

故事线缘起于2009年R. Munroe开创的XKCD手绘插图“电影叙事”，漫画通过线条叙述方式由左至右展示人物角色，以线条的靠近或偏离代表人物交互

* 本文系国家社会科学基金项目“数字人文视角下历史档案资源知识聚合与知识发现研究”（项目编号：19BTQ102）研究成果之一。
作者简介：王阮，助理研究员，吉林大学鼎新学者，博士后；邓君，教授，中国人民大学档案事业发展研究中心研究员，博士，博士生导师，通信作者，E-mail：dengjun9722@163.com。
收稿日期：2021-10-07 修回日期：2021-11-28 本文起止页码：17-25 本文责任编辑：徐健

会话的开始与结束[8]，不仅能显示事件在时间上的先后顺序，还可以表示事件间的语义关系以及衍生事件，直观地展示事件随时间发展和传播的过程[9]。

起初，受电影漫画研究启发，故事线以数字叙事或虚拟叙事、交互式叙事的形式出现[10]，国内外学者们致力于探寻故事线组成要素。M. Bal 将故事线元素划分为事件、角色、时间、地点以及其他（情节、语气、观点）[11]，T. Tan 等在 Bal 分类基础上将元素重组为角色、关系、结构、修饰、情节和事件五类[12]，佘玉轩等认为故事线可视作日期、时间、机构、人物、地点、主题和关键词的联合概率分布[13]。

伴随研究不断深入发展，故事线衍生为可视化叙事[14]、数据驱动叙事[15]等不同概念形式，学者们专注于故事线构建方式及提升故事线条美观度。如在 Lucene 检索结果集上构建多视点图，随后通过寻找最小权重支配集来筛选代表性数据，最后通过求解有向斯坦纳树问题生成故事线[16]；或基于贝叶斯网络无监督挖掘算法实现故事线自动布局[17]；亦或基于遗传算法（genetic algorithm，GA）计算故事线布局策略，引入"交互对话"概念，应用初始会话布局、线条优化等提升故事线构建效果[18]。同时，因 GA 算法所属人工智能算法，故而布局所消耗时间较长，因此需要借助平滑算法调整线的几何形状[19]或应用 StoryFlow、StoryCake、Yarn、LitStoryTeller、TimeNets 等工具或工具组合使用[20]来优化故事线呈现形态，从而使故事线条更加清晰。

受数字技术影响，学者们不再拘泥于探略故事线构建形态，而是希冀深入挖掘要素、分解要素，使其精细化、碎片化，进而整合要素、串联要素，实现可视化、情感化、场景化展示。相关研究主要集中在计算机领域，如通过改进单源最短路径发掘话单数据，以 Spark Graphx 和 Echarts 实现相关特征和人物关系图的可视化[21]。利用视频流分解算法、关键帧提取算法对视频段分解、抽取情节，实现故事主线的时序结构展现[22]。同时在大量数据梳理及挖掘过程中，按照数据处理粒度与维度不同，抽取情感特征语词，获取情感持有者意见[23]，掌握情感波动轨迹。除此之外，图书情报领域也对此类问题展开了相关研究。陈博等基于文本挖掘技术提取《英雄格萨尔》主题特征词，实现故事主题可视化[24]；欧阳剑采用可视化分析方法对大规模古籍文本进行挖掘，不断改进与优化应用场景的分析发现[25]。作为一种流行的策划、组织和个人叙事的方法[26]，王晰巍等从故事线、时间线、情感线等方面对社交网络事件进行知识图谱可视化分析[27]；X. M. Zou 等结合社会场景和主题情境梳理微博故事情感线，引入话题上下文模拟语义关系[28]；张海涛等根据 Louvain 算法划分评论网络群落，动态跟踪热点事件网民话题意见并抽取情感特征词，实现舆情故事线场景化与情感化[29]。K. Mcdowell 认为故事线是通过叙事经验构成的，并定义了一个信息框架即数据、信息、知识、智慧框架（DIKW）以展示故事线和故事叙事应该如何引发概念范式转变[30]。

综上所述，学者们从探寻故事线要素，过渡到应用相关工具、算法构建故事线与优化故事线形态，再到窥探要素的呈现。然而在现有研究中，故事线要素展示较为局限，仅仅围绕 1 个（如事件）、2 个（如情感和时间或情感和事件）要素展开研讨，较为侧重事件相似性分析，易忽略故事线完整的结构化表达，无法直观洞察关联要素（如人物、时间、地点、事件、情感等），实现多维要素总览，并且辐射范围以计算机通信、社交媒体、网络舆情等领域为主，鲜从历史人物挖掘视角进行探究。数字时代，故事线的相关研究仍处于不断探索与发展阶段。在数字浪潮推动下，人文研究者们亟需新的技术工具和人文逻辑相耦合的新研究范式，亟待技术的"生鲜注入"，拓宽故事线研究视角，拓展数字人文跨域空间，推动实践纵深发展。

因此，基于数字人文视角，本文将"数字技术"嵌入"人文研究"，以《张学良口述历史》文本为研究对象，引入 LDA 主题模型与 SnowNLP 情感词典，以 jupyter 为实验工具，借助 jieba 分词对文本进行预处理，识别人物、地点、事件、时间、情感等关联要素，构建多维可视、情感集成的张学良故事线，有助于人文学者读史、学史、品史、鉴史、思史，实现文本挖掘创新，同时提供了基于文本数据的故事线构建及可视化、情感化、场景化全新研究范式。

## 3 研究框架及研究方法

### 3.1 研究框架

历史记载最常见的史料形态是文本，伴随数字人文时代到来，"口述"形式突破了文本叙事牢笼，铸就了文本、图片、音频、视频等多元并包的新业态，丰富了历史档案研究。同时，口述历史是史料呈现最直观化、写实化的表示，能以颇具生动性的"口述"叙事方式补充、还原史实。

在本研究中，首先对初始数据源进行预处理，获取实验数据源。然后应用 LDA 主题识别模型获取主题

分布，进行 t-SNE 数据降维并呈现主题模块，同时借助 Python 类下 SnowNLP 情感词典逐行对文本数据进行情感挖掘，将情感演化过程予以可视化呈现，以此为基点梳理张学良故事线，研究框架如图 1 所示：

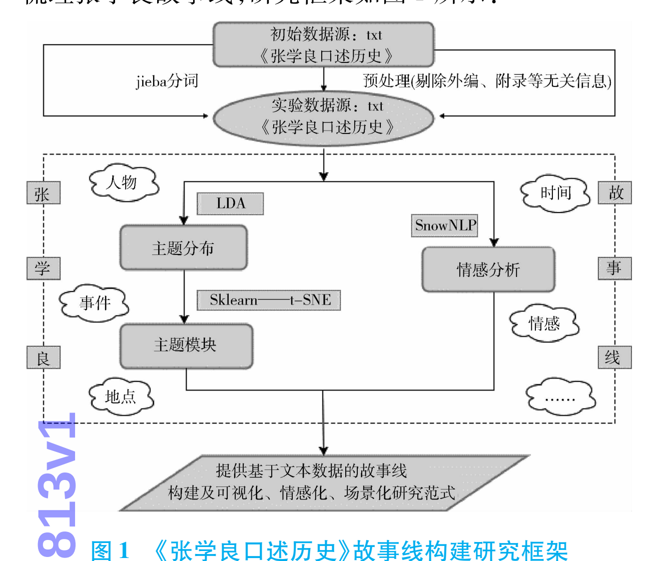



图 1 《张学良口述历史》故事线构建研究框架

## 3.2 研究方法

### 3.2.1 LDA 主题模型

隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）是一种文档主题生成模型，也称为三层贝叶斯概率模型，最早由 D. M. Blei 等[31]提出，可将文档词向量约简为主题时空降维表达，在处理文本过程中具备较高的模型泛化能力、良好的灵活性与适应性，具有文本挖掘与聚类、主题划分与解析等多样化功能。

本研究将 LDA 主题建模方法应用于文本内容抽取，可实现文本语义层面向多元主题空间聚类过渡，从而辅助相关研究者挖掘所需有用信息，实现精准感知与智能抽取。值得注意的是，在主题识别过程中，LDA 模型主题数目将影响主题识别效果[32－33]。主题数目设置过多，会造成识别出的主题分布过于稀疏，主题相似度过高；主题数目设置过少，会导致主题过于宽泛，无法准确揭示文本核心内容[34]。而困惑度（perplexity）可以用来确定合适的主题数目，度量概率分布或概率预测样本的好坏程度。

### 3.2.2 SnowNLP 情感词典

文本情感分析（text sentiment analysis）又称为意见挖掘，是对带有情感色彩的主观性文本进行分析、处理、归纳和推理[35]。按照文本粒度，可分为词语级、短语级、语句级、篇章级以及多篇章级等研究梯次[36]。目前，常见的文本情感分析有两种路径：基于情感词典和基于机器学习。本研究采用基于 python 的 SnowNLP 情感词典分析方法，自带训练好的基础情感词典和自建词典。

鉴于在实际操作过程中，情感词典的完备性会直接影响情感特征词的提取效果，并进一步影响最终的实验结果。因此，在实验前，笔者首先通过抽取部分文本数据导入基础情感词典测试，发现情感语词结果所属类别较为规范，准确性较高，实验效果良好。如“不大正直”“非常艰苦”“不喜欢”“太不争气”等所属“消极”，“聪明”“很好”“漂亮”“年轻”等所属“积极”，故而经测试后，无需自建词典，选取基础情感词典即可。

# 4 数据采集和处理

## 4.1 数据来源及预处理

本文爬梳西安事变数据库《张学良口述历史》文本作为初始数据源（txt）[37]，剔除外编、附录等无关信息后进行 jieba 分词处理，实验工具为 jupyter notebook。部分分词结果如图 2 所示。由此获取实验数据源，以 unicode utf-8 编码，txt 格式保存，为梳理张学良故事线脉络提供可靠数据支撑。

出版 说明
翻开 中国 当代史 张学良 近百年来 影响 中国 历史进程 人物
过去 一百年 中国 出现 翻天覆地 变化 风云人物 涌现 真正 改变 中国 发展 轨迹 改写 中国 历史 人物 屈指可数 张学良 先后 两度 历史 关键时刻 国家 统一 抗日救亡 大是大非 问题
张学良 有着 传奇 一生 活动 多次 改变 历史 方向 皇姑屯 事变 后 不久 张学良 宣布 东北 易帜 统一 国民政府 西安事变 促使 蒋介石 抗日 掀起 光辉灿烂 一页 张学良 杨虎城 将军
张学良 将军 2001 年 10 月 14 日 美国 夏威夷 与世长辞 享年 101 岁 时任 中共中央 总书记 国家 主席 江泽民 发 去 唁电 高度评价 张学良 历史 功绩 誉其为 中华民族 千古 功臣
张学良 口述 历史 系 缘于 哥伦比亚大学 哲学 教授 史学家 唐德刚 博士 1990 年 1 月 5 月间 台北 北投 张学良 寓所 亚 饭店 先后 录下 11 盘 录音带 录音带 标注 录音 时间 分别
张学良 看 唐德刚 撰写 李宗仁 回忆录 后 派 人 找到 唐德刚 先生 说 张学良 想 请 吃饭 那次 宴会 张学良 表示 想 请 唐为 写 回忆录 后来 种种原因 没有 完成 全部 预定 工作
唐德刚 深有感触 地说 海外 华裔 史学工作者 眼底 手头 见 琳琅满目 中华 无价之宝 眼睁睁 看着 逐渐 流失 内心 发生 沉重 使命感 遗恨 惋惜 之情 交织 无能为力 心理 孤独 之感
张学良 将军 辞世 后 唐德刚 先生 助手 帮助 下 历经 数年 精心 整理 录音 资料 成本 书 全书 张学良 自述 内容 再现 张学良 精彩绝伦 一生 作者 忠实 张学良 自述 历史 记述 全部
该书 出版 大众 读者 历史 研究者 很强 现实意义 历史 意义
张学良 自述 是是非非 代序

图 2 《张学良口述历史》文本 jieba 分词部分展示

## 4.2 结果分析

### 4.2.1 主题模型可视化

文本主题可视化表示是指把文本知识转化为用图形、图像或动画表示的知识，其目的在于让人直观地观察到核心信息和关键数据，从而快速发现其中蕴含的深层知识[38]。透过主题模块，用户可以看出特定时间点上主题的分布，以及多个主题随时间的发展变化。LDA 模型采用词袋（bag of words）方法，将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息，简化了主题分析的复杂性。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

(1)LDA 主题模型识别。本文将 jieba 分词后的 txt 文本存储读取语料,通常每一行文本可视作一篇文档,生成 Doc0-Doc1015,共计 1 016 条文本。此时文本的词语转换为词频矩阵,矩阵元素 a[i][j] 表示 j 词在 i 类文本下的词频,根据词频矩阵,应用 LDA 算法,迭代 500 次,提取关键特征,进一步获取文档主题分布,输出主题中的 TopN 关键词、主题个数。同时,以困惑度作为衡量本研究中《张学良口述历史》文本主题数目划分是否科学的依据。通常来说,低困惑度能更好地预测样本主题数目,即困惑度越低,聚类效果越好。困惑度值如图 3 所示,主题数目为 5 时,困惑度达到最低值。因此,《张学良口述历史》文本主题划分数目是 5 时聚类效果最佳。

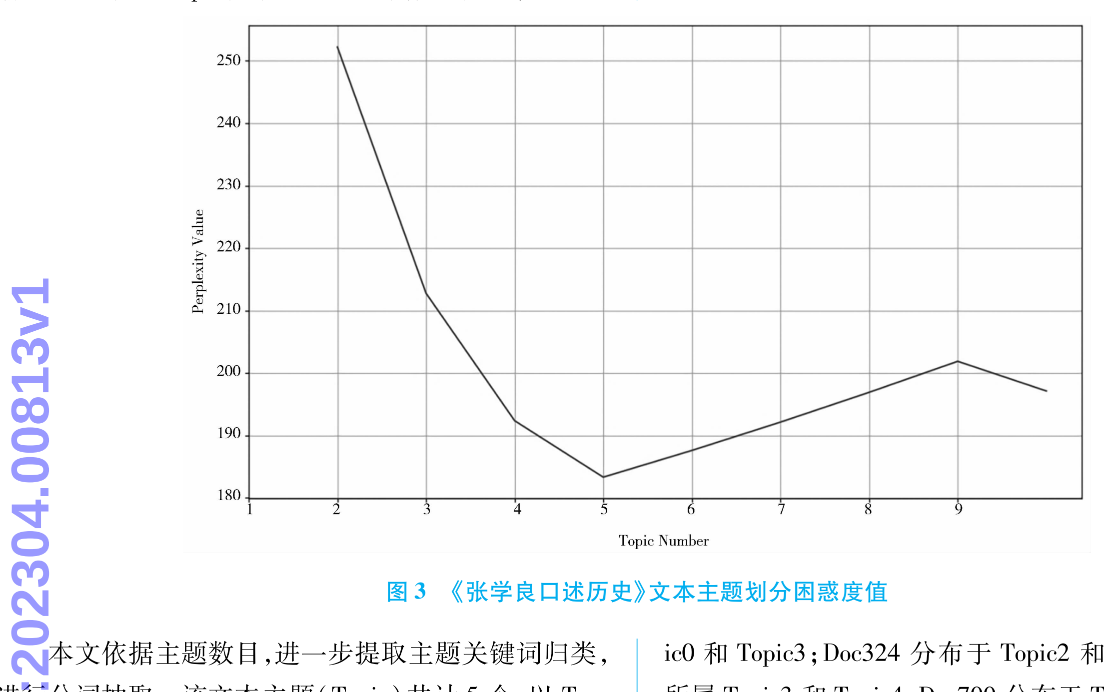



图 3 《张学良口述历史》文本主题划分困惑度值

本文依据主题数目,进一步提取主题关键词归类,进行分词抽取。该文本主题(Topic)共计 5 个,以 Topic0-Topic4 分布,结合主题分布结果和《张学良口述历史》文本信息,概括主题如下:即 Topic0——身世背景:从张家发迹入手,主要介绍了张学良祖母,爷爷奶奶,父母,兄弟姐妹,姑姑姑父,二伯父等家世背景。Topic1——女人韵事:主要讲述张学良与妻子于凤至、谷瑞玉、赵一荻(赵四),以及梁九、墨索里尼小姐(墨索里尼的女儿)等众多女性朋友的情感纠葛。Topic2——幼青时代:即描述张学良与张作霖的父子之情以及张学良初入讲武堂弃文从军之事。Topic3——少帅之路:以叙述张学良带兵之道为背景,围绕吴佩孚、郭松龄、冯玉祥等人展开叙事,以东北易帜、热河失守、皇姑屯等东北往事为主线展开,涉及第一次直奉战争、第二次直奉战争、南口军纪案等。Topic4——晚年生活:以张学良晚年生活为主,如子孙现状、个人喜好等。

与此同时,调用 matplotlib. pyplot 输出文档对应的主题分布图。笔者随机抽取文档 Doc25、Doc166、Doc288、Doc324、Doc501、Doc700,查看并验证文本所属主题领域,如图 4 所示。由此可见,Doc25 所属 Topic2;Doc166 分布于 Topic1 和 Topic3;Doc288 所属 Topic0 和 Topic3;Doc324 分布于 Topic2 和 Topic4,Doc501 所属 Topic3 和 Topic4;Doc700 分布于 Topic4。

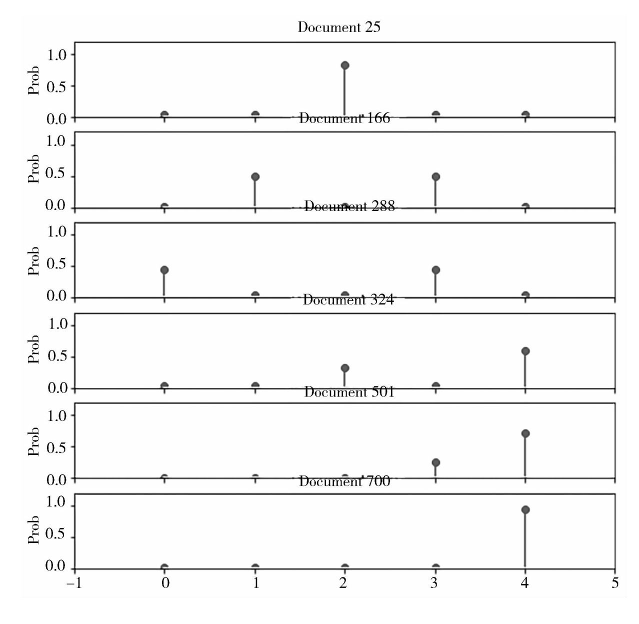



图 4 随机文档所属主题分布

(2)t-SNE 降维。在高维数据中,通常用多特征表示一个数据点,这不仅为数据描述和分析带来诸多困难,同时也会增大该数据点的计算难度和计算所需空

间和时间。因此,为避免数据训练时间过长,为使上述主题模块更为清晰呈现,提升数据可视化效果,本研究采用 t-SNE 降维。

t-SNE(t－分布式随机邻域嵌入)作为挖掘高维数据的非线性降维算法[39],通过多特征数据点的相似性来识别、观察,进一步发现数据规律,实现多维数据向 2 维或 3 维空间映射。经 t-SNE 降维后,Topic0-Topic4 以团簇形式呈现,5 个主题(Topic0-Topic4)以 0－4 标号分布,如图 5 所示:

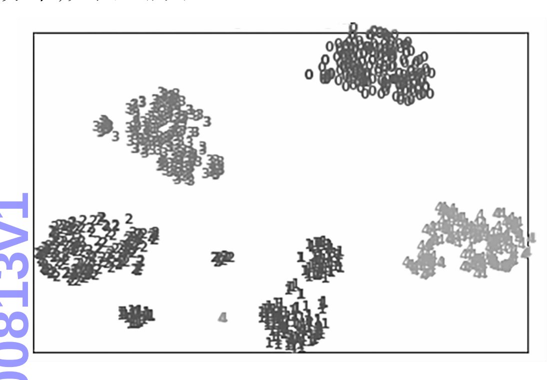


图 5　t-SNE 主题模块

### 4.2.2　情感态度可视化表示

情感态度可视化表示是指以图信息呈现文本蕴含的情感态度,实现情感走向的可视化映射。文本除了包含主题、观点和结构之外,还蕴含肯定、否定、喜爱、厌恶、赞赏、批评等情感态度信息[30]。伴随事态发展,口述者通常在不同时期会发生情感波动,即情感变化,因而,抽取情感特征词进行情感极性分析,绘制动态变化的情感演化图谱,能直观呈现张学良在各个时期的情感动态演化。选择 SnowNLP 基础情感词典对实验数据源进行情感分析,加载 sentiment 情感分析模块对语料库进行训练,同时调用 matplotlib 输出情感波动结果,如图 6 所示:

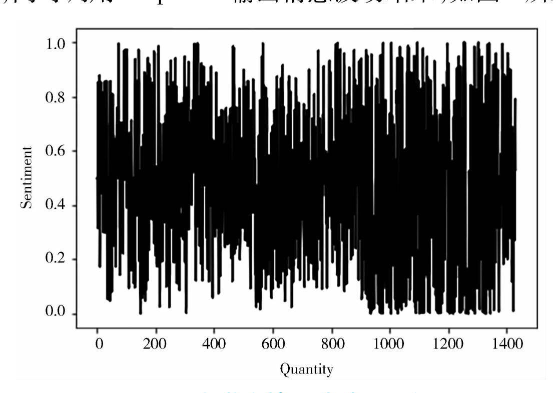



图 6　张学良情感波动图示例

从图 6 看出,每一条文本信息的 sentiment 取值以线条接近取值边框距离为基准,sentiment 取值接近 1 为积极,表示情感倾向正面情绪;sentiment 取值接近 0 为消极,表示情感倾向负面情绪。就张学良而言,其情感线历经情感积极期—情感中立期—情感消极期—情感平淡期 4 个阶段。

前 400 条文本信息 sentiment 接近 1,所属情感积极期,以"聪慧""一表人才""漂亮""阔气""赞成""很好""非常高兴""得意""愿意""贤妻良母""恩爱""非常喜欢"等情感语词为主,该阶段故事线以张学良身世背景、父子关系、求学经历、婚姻故事为主线展开;400－800 条文本信息 sentiment 介于 0 和 1 之间,所属情感中性期,主要情感语词涉及"说不定""复杂""奇怪""差不多""舍不得""踌躇"等,该阶段故事线以战争背景为主旋律,且涵盖第一、二次直奉战争,北京政变。既富含浓烈的爱国主义情怀,又饱含对叛变倒戈之人的愤怒;800－1 400 条文本信息 sentiment 接近 0,所属情感消极期,且消极语词密度略显密集,以"恨透了""忏悔""自责""痛苦""厌烦""悲愤""谴责""羞辱""难过"等情感语词为代表,该阶段以内战、"九一八"事变、"西安"事变等重大转折性历史事件为主,既有失去国土领地的心痛,又有国共内战的厌倦之情,同时又兼有遭受"忏悔录"风波的误解与迷惑。1 400 条文本信息所属情感平淡期,此段经历主要介绍张学良晚年生活,偶有消极情感波动。

### 4.3　张学良故事线可视化、情感化、场景化呈现

故事线旨在将繁琐复杂的数据进行图标化的整合,从提炼分析出的数据组织故事[3],由关联要素(人物、事件、时间、地点、情感等)组建,关联要素亦可称之为故事单元(story unit),每个故事单元又是发生在某个特定的场景中[15],而单个故事单元仅能反映故事线的单一层面,无法揭示故事线完整全貌。故而,本文试图将零散的故事单元有机集成,通过汇集诸多故事单元完善张学良故事线完整表达。

故事线可视化即重组展示数据形态,为进一步场景化建构提供必要基础。故事线情感化即将情感线条渗入故事情境,进而丰富场景丰聚度。故事线场景化即由粗粒度过渡到细粒度的组织过程,塑造人、事、地、时、情感等要素的"集合圈"。可视化、情感化、场景化相互交织互促以完善张学良故事线条,成为推动叙事内容更加丰富的重要手段。

基于前述主题模块分布及情感线梳理,笔者在 jupyter 实验工具中读取人物要素、事件要素、时间要素、地点要素,并以时间线为横轴,串联附加情感演化各个时期,同时读取文本大事年表,对张学良故事线发展脉络进行梳理,如图 7 所示:

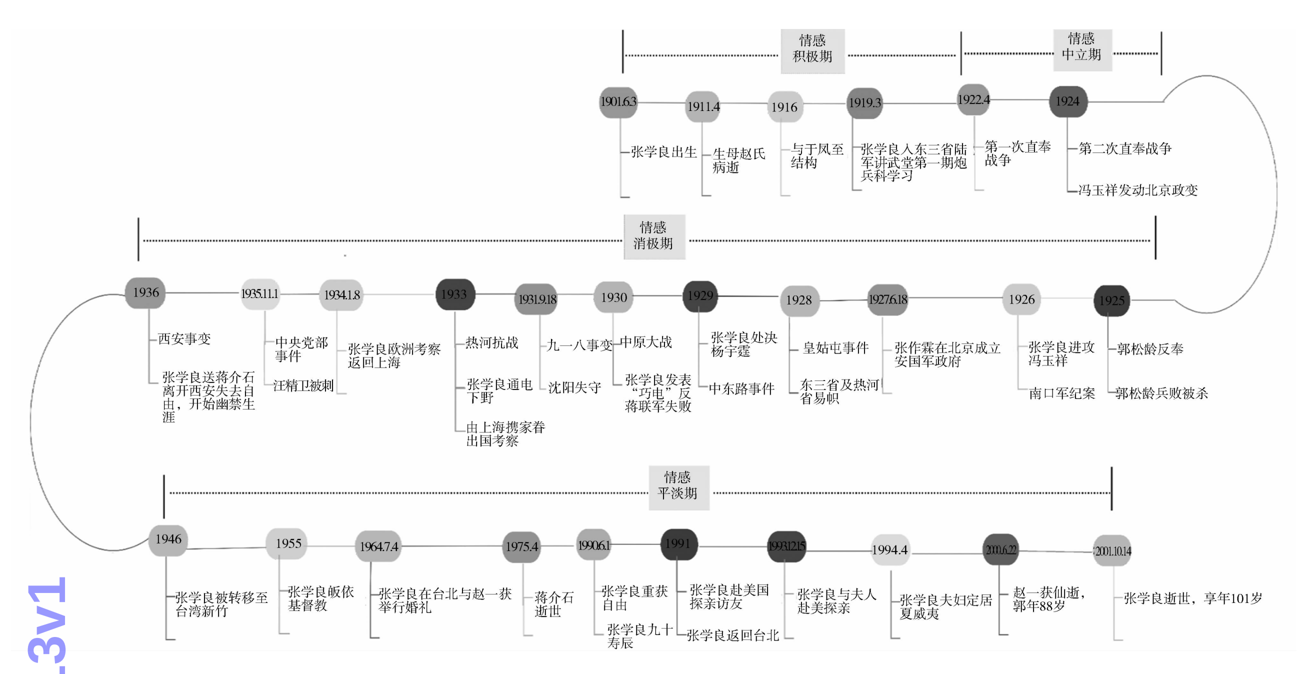



图 7　张学良故事线及可视化、情感化、场景化呈现

总体而言，聚焦“人”的故事维度。张学良故事线存在大量以人物为中心的叙事素材，如赵氏、于凤至、冯玉祥、蒋介石、杨宇霆、汪精卫等，这些人物是深入了解张学良的重要窗口。聚焦“事”的故事维度。事件有助于进一步回溯揭示真相，还原史实内容。聚焦“地”的故事维度。地点凸显着地域布局变化，彰显着地域形态迁移之势，有利于探究人物移动轨迹。聚焦“时”的故事维度。以时间轴将时间节点联结，清晰可见张学良一生的时间跨度，有助于准确、详实地把握人物生平脉络。

细化而言，从图 7 可以看出，1901 – 1924 年展示了张学良情感演化两个阶段。情感积极期着眼于张学良少时的成长经历，涵盖 4 个时间切片：1901 年、1911 年、1916 年和 1919 年。人物涉及其生母赵氏，妻子于凤至；情感中立期包含 1922 年和 1924 年两个时间切片，记载了第一次、第二次直奉战争，涉及的人物是冯玉祥。1925 – 1936 年囊括了 11 个时间切片，涉及的人物有汪精卫、蒋介石、杨宇霆、郭松龄、张作霖、冯玉祥，呈现了“西安”事变、中央党部事件、热河抗战、“九一八”事变、皇姑屯事件、中原大战等事件，此阶段张学良情感呈现消极状态。情感平淡期（1946 – 2001 年）以张学良晚年生活为主线，包含 10 个时间切片。以 1991 年为例，该时间节点记录了张学良携夫人赴美国探亲访友，而后返回台北，展示了地点变化。

由此可见，透过线条梳理，清晰可见张学良完整故事线（自 1901 年 6 月 3 日出生到 2001 年 10 月 14 日逝世），包含与之关联的人物、事件、地点、时间，同时予以情感演化各阶段呈现，促进了文本单一记事转向多维叙事，实现了关联要素可视化、情感化、场景化动态互促。

## 5　讨论与启示

传统的文本叙事需要借助人工方式手动抽取相关知识内容。而数字人文背景下，数字技术的加持能拓展文本范型的记事表现力，切实解决传统人工模式难以快速捕捉关联知识的困囿，彰显数字人文视野下的巨观数据计算优势，展现知识要素从文本记事向多维叙事场域迁移，实现数字研究与人机共读，凸显“数字”与“人文”内核。

### 5.1　从文本史料到故事多维呈现

传统物理媒介是静态的、被动的，而数字媒介则是动态的、能动的。故事线的表现形式使得人文研究不仅仅显现于文本史料表象，而是透过“线条”较为精细化、细粒度描绘，成功将与张学良关联的人物、时间、事件、地点、情感等文本信息与引人入胜的故事情节串联，将过往历史予以揭示，辅助人文学者提炼与挖掘生动有趣的故事素材，快速定位查找史料对应的“关联数据”继而进一步理顺历史“脉络”。同时，因“口述”研究是从原始史料中提取故事要素，展现了知识素材的原始状态，故而从计算软件中提取丰富多元的人物信

息、逻辑清晰的故事线索、跌宕有致的事件过程创设故事线,不仅实现了文本信息传递与“活态”史料互动,而且通过穿插在历史叙事中的语义描述,激发了人文学者进行深层次知识发现、知识迁移、知识挖掘与知识创新。

### 5.2 从文本表达到视觉可视化表征

故事线条的生动展现实现了文本“知识表示”向可视化“知识单元”过渡,避免了史料潜藏于传统书籍的叙述,将零散杂乱、可读欠佳、检索不便的口述历史文本集汇,经数字技术解构、重组、重构后更易于梳理、理解与应用,催生了极具冲击力与感染力的活态数据可视化集成。一方面,通过“引介”故事线拓宽了文本知识表示的可视化路径;另一方面,注入新技术方法抽取关联信息,集聚了可视化要素特征,如主题识别、时间序列分析、人物关系探索,以此来进行事件史实揭示与阐释、历史发展规律与趋势解读等,加速文本数据向故事线图谱聚类表示的知识场域迁移。同时,故事线条可视化视觉表征除了传达文本事实,其目标还在于传输态度、期望、见解等,并以此匡扶人文研究者正确解构、重组、思辨知识,藉由可视化故事线条“关联数据”,构筑知识关联网系。

### 5.3 从文本语词到情感场域集汇

情感分析对于透析人物属性至关重要。情感多源于现实生活,故而以情感语词作为情感判别依据,有助于真实还原人物形象,将情感演化过程融入故事线,有助于厘清张学良伴随事态变迁的情感极性与心路历程。我们应当注意到,在张学良情感动态演化过程中,当情感伴随故事发展产生波动时,有时不仅仅表现为单一情感作用,而是多元情感交错互替。张学良在不同时期不同阶段的经历致使其情感极具波动性,衍生出积极期—中立期—消极期—平淡期 4 个阶段,既折射出张学良浓烈的爱国主义情怀,收复东北、匡复河山的雄心壮志,又凸显其深感己力和国力不足的遗憾,无法报国恨家仇的愤慨以及“仇日情结”的不断升级,同时还伴有“忏悔录”风波的疑惑,以及晚年处事的淡然安逸等,多元情感共同集汇于情感场域,贯穿张学良整个故事线始终。

### 5.4 从文本叙事到场景重构

以“线条”清晰描绘故事脉络,而非以传统人工手动逐页浏览阅读书籍,突破了扁平化的单调文字叙述,让数据表现形式更加“鲜活”,实现了从文本叙事到数据解析,再到多维场景重构。透过故事线,能清晰可见与张学良关联的人物(赵氏、张作霖、于凤至、赵一荻、郭松龄、冯玉祥、杨宇霆、蒋介石、汪精卫等)、事件(第一次直奉战争、第二次直奉战争、“九一八”事变、“西安”事变等)、地点(北京、西安等)、时间序列(1901.6.3、1911.4、1916 等)、情感演化时期(积极—中立—消极—平淡)等。

## 6 结论与展望

通过抽取机读后可识别的知识要素构筑张学良故事线,实现了化繁为简,铸就了更加清晰、直观的人—事—地—时—情感一体化故事线脉络体系。理论层面,本文为人文领域人物脉络梳理、故事线构建及可视化、情感化、场景化实现创新了研究范式,丰富了框架构想。实践层面,以《张学良口述历史》文本为数据源进行应用探索,为进一步剖析张学良传奇人生提供了一定指导。同时,透过故事线条视觉表征,提供了交互式历史数据展示,可辅助人文研究者快读、远读,迅速捕获张学良一生重要时间点、重要事件、主要地点、关联人物以及情感演化变化,有利于进一步洞察与深入挖掘新问题,拓宽知识研究视野,推动知识价值增益,助力“数字”与“人文”互融共生。

本文在研究过程中仍然存在一定局限性。如对《张学良口述历史》文本的主题内容可视化呈现仅采用 LDA 单一算法,主题方法的选择以及是否可以应用多元算法更加全面化、细粒度与精准化提取知识要素,选择不同情感词典进行情感分析,使数据维度不再局限于扁平化处理,实现图谱聚类、3D 多样化、地理时空化、VR/AR/MR 虚拟化等在未来研究中可进一步探索与补充。相信未来“数字”与“人文”的互促与融合会更加生动有趣,数字技术的助力也必将加速人文知识开掘。

**参考文献:**

[1] 唐家渝,刘知远,孙茂松. 文本可视化研究综述[J]. 计算机辅助设计与图形学学报,2013,25(3):273－285.

[2] 刘炜,叶鹰. 数字人文的技术体系与理论结构探讨[J]. 中国图书馆学报,2017,43(5):32－41.

[3] 朝乐门,张晨. 数据故事化:从数据感知到数据认知[J]. 中国图书馆学报,2019,45(5):61－78.

[4] CHEN X. Why did John Herschel fail to understand polarization? The differences between object and event concepts[J]. Studies in history and philosophy of science,2003,34(3):491－513.

[5] 朝乐门. 数据科学理论与实践[M]. 北京:清华大学出版社,2017.

[6] 丁家友,唐馨雨. 数字人文视角下的数据叙事及其应用研究[J/OL]. 情报理论与实践:1 - 11[2021 - 11 - 24]. http://kns.cnki.net/kcms/detail/11.1762.G3.20210903.1446.004.html.
[7] 伯迪克, 德鲁克, 伦恩费尔德,等. 数字人文:改变知识创新与分享的游戏规则[M]. 马林青,韩若画,译. 北京:中国人民大学出版社,2018:51.
[8] MUNROE R. XKCD[EB/OL]. [2021 - 10 - 30]. https://xkcd.com/#.
[9] NAZANIN D, MASOUD A. SGSG: semantic graph-based storyline generation in Twitter[J]. Journal of information science, 2019,45(3):304 - 321.
[10] BALET O, SUBSOL, G, TORGUET P. Virtual storytelling. using virtual reality technologies for storytelling [M]. Cham: Springer,2001.
[11] BAL M. Story, text, and scripture-literary interests in biblical narrative-KORT, WA [J]. Theology today, 1989, 45 (4): 460 - 464.
[12] TANG T, RUBAB S, LAI J W, et al. iStoryline: effective convergence to hand-drawn storylines[J]. IEEE transactions on visualization & computer graphics, 2019,25(1):769 - 778.
[13] 余玉轩,熊赟. 基于贝叶斯网络的故事线挖掘算法[J]. 计算机工程,2018,44(3):55 - 59.
[14] SU J, DAI Q, GUERIN F, et al. BERT-hLSTMs: BERT and Hierarchical LSTMs for visual storytelling[J]. Computer speech & language, 2021,67:101169 - 101182.
[15] STAMATIADOU M E, THOIDIS I, VRYZAS N, et al. Semantic crowdsourcing of soundscapes heritage: a mojo model for data-driven storytelling[J]. Sustainability, 2021, 13(5):1 - 19.
[16] 李培,翁伟,林琛. 中文微博故事线生成方法[J]. 中文信息学报,2016,30(3):143 - 151.
[17] OGAWA M, MA K L. Software evolution storylines[C]// Proceedings of the ACM 2010 symposium on software visualization. New York: ACM, 2010: 35 - 42.
[18] TANAHASHI Y, Ma K L. Design considerations for optimizing storyline visualizations[J]. IEEE transactions on visualization & computer graphics, 2012, 18(12):2679 - 2688.
[19] 彭燕妮,樊晓平,赵颖,等. 时间事件序列数据可视化综述[J]. 计算机辅助设计与图形学学报,2019,31(10):1698 - 1710.
[20] LIU S, WU Y, WEI E, et al. Story flow: tracking the evolution of stories[J]. IEEE transactions on visualization & computer graphics, 2013, 19(12):2436 - 2445.
[21] 毛辰阳. 基于 Spark 平台及话单分析的人物关系可视化的研究与应用[D]. 北京:北京工业大学,2018.
[22] 王东辉,朱森良,吴春明. 基于时序结构图的视频流描述方法[J]. 计算机学报,2001(9):944 - 950.
[23] FELDMAN R. Techniques and applications for sentiment analysis [J]. Communications of the ACM, 2013, 56(4):82 - 89.
[24] 陈博,陈建龙. 基于文本挖掘和可视化技术的主题自动标引方法——以《英雄格萨尔》为例[J]. 现代情报,2019,39(8):45 - 51,102.
[25] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016,42(2):66 - 80.
[26] COPELAND S, DE MOOR A. Community digital storytelling for collective intelligence: towards a storytelling cycle of trust[J]. AI & society,2017,33(11):101 - 111.
[27] 王晰巍,贾若男,韦雅楠,等. 社交网络舆情事件主题图谱构建及可视化研究——以校园突发事件话题为例[J]. 情报理论与实践,2020,43(3):17 - 23.
[28] ZOU X M, YANG J, ZHANG J P, et al. Microblog sentiment analysis using social and topic context [J]. Plos one, 2018, 13(2):e0191163.
[29] 张海涛,刘雅姝,张枭慧,等. 基于模块度的话题发现及网民情感波动研究——以新浪微博"中美间贸易摩擦"话题为例[J]. 图书情报工作,2019,63(4):6 - 14.
[30] MCDOWELL K. Storytelling wisdom: story, information, and DIKW[J]. Journal of the Association for Information Science and Technology, 2021,72(10):1223 - 1233.
[31] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5): 993 - 1022.
[32] PONWEISER M, GRUN B. Finding scientific topics revisited [C]//Advances in latent variables. Berlin: Springer, 2014: 93 - 100.
[33] 关鹏,王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术,2016(9):42 - 50.
[34] 蒋甜,刘小平,刘会洲. 基于关键词关联度指标(KRI)进行 LDA 噪声主题过滤的方法研究[J]. 图书情报工作,2020,64(3):92 - 99.
[35] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations & trends in information retrieval, 2008, 2(1/2):130 - 135.
[36] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报,2010,21(8):1834 - 1848.
[37] 山西省图书馆. 西安事变数据库[EB/OL]. [2021 - 10 - 30]. http://www.sxlib.org.cn/dfzy/xasb/rwz/zxrw_18592/zxl_18593/zxlcksj_18595/zxlksls_18606/.
[38] 马创新,陈小荷. 文本的可视化知识表示[J]. 情报科学,2017,35(3):122 - 127.
[39] KOHLER M, SONDERMANN L, FORERO L, et al. Classifying and grouping narratives with convolutional neural networks, PCA and t - SNE[C]//Proceedings of the hybrid intelligent systems. Cham: Springer, 2020: 22 - 30.

**作者贡献说明:**

王阮:论文撰写与修改;
邓君:论文审阅与写作指导。

## Storyline Construction and Application Exploration of Visualization, Emotion and Scene: Taking *Zhang Xueliang's Oral History* as an Example

Wang Ruan　Deng Jun
School of Business and Management, Jilin University, Changchun 130012

Abstract: [Purpose/Significance] Combing the historical facts with storyline not only has a certain theoretical and practical guidance and significance for describing and grasping the direction of historical development, but also provides a new technology realization mode and innovative research perspective for knowledge discovery in the humanities field. [Method/Process] This study provided a research paradigm of storyline construction and visualization, emotion and scene based on text data, and made empirical exploration by combing the text of *Zhang Xueliang's Oral History* as the data source. This paper eused jieba word segmentation to clean the initial data source of *Zhang Xueliang's Oral History* to obtain experimental data source. LDA topic model was used to obtain topic distribution and t-SNE data dimension reduction was performed to present topic module. With the help of SnowNLP emotion dictionary, emotional feature words were mined, Zhang Xueliang's emotional evolution stage was sorted out, and the storyline was constructed. [Result/Conclusion] Through the construction of Zhang Xueliang's storyline, the dynamic mutual promotion of multi-dimensional elements such as characters, places, events, time and emotions is realized.

Keywords: storyline　visualization　emotion　scene　Zhang Xueliang



### 中国科学院科研道德委员会办公室发布关于规范学术论著署名问题负面行为清单的通知

近日，中国科学院科研道德委员会办公室发布《关于规范学术论著署名问题负面行为清单的通知》（科发监审函字〔2022〕1 号），对中国科学院学术论著署名问题进行了规范要求，并列出了学术论著署名问题的负面行为清单。

通知指出，科研诚信是科技创新的基石。维护科研诚信、开展负责任创新，既是中国科学院科研人员从事科学研究、推进科技创新的基本原则，也是其作为国家战略科技力量主力军定位的基本要求。学术论著署名规范一般由学术界长期形成的惯例自行确定，根据学科、领域甚至不同的科技期刊均可能有不同的规范要求。制定出适用于不同场景的统一署名规范较为困难。通知列出了部分学术论著署名问题的负面行为清单，如冒用作者署名、虚构作者署名；无实质性贡献的人员参与署名；未经所有作者一致同意就确定署名顺序；违反署名第一作者或通讯作者时的必要性原则而罗列过多的第一作者或通讯作者；因作者所属机构变化而随意变更论著工作主要完成机构；虚构、伪造作者所属机构；把论著非完成机构作为署名单位；使用非正式联系方式作为论著作者的联系方式；故意排斥有重要贡献的科研工作者参与署名；侵害直接实施科学实验的研究生的基本署名权等。

中国科学院对清单所列行为实施“零容忍”要求，要求凡中国科学院科研人员出现清单所列行为，将由相应第一责任单位按照通知的相关规定开展调查，并根据具体事实和相关情节予以认定和处理，对严重违背科研诚信要求的行为终身追责。

（本刊讯）